BANK OF CANADA
BANQUE DU CANADA

Working Paper/Document de travail
2014-51

Bootstrap Tests of Mean-Variance Efficiency with Multiple Portfolio Groupings

by Sermin Gungor and Richard Luger

www.bank-banque-canada.ca

Bank of Canada Working Paper 2014-51

November 2014

Bootstrap Tests of Mean-Variance Efficiency with Multiple Portfolio Groupings

by

Sermin Gungor[1] and Richard Luger[2]

[1]Financial Markets Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
sgungor@bankofcanada.ca

[2]Département de finance, assurance et immobilier
Université Laval
Quebec, QC, Canada
Richard.Luger@fsa.ulaval.ca

Bank of Canada working papers are theoretical or empirical works-in-progress on subjects in economics and finance. The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada.

ISSN 1701-9397

© 2014 Bank of Canada

Acknowledgements

We would like to thank Ian Christensen, Antonio Diez de los Rios, Jonathan Witmer, Bo Young Chang and seminar participants at the Bank of Canada for helpful comments and useful conversations. All remaining errors and omissions are our own.

Abstract

We propose double bootstrap methods to test the mean-variance efficiency hypothesis when multiple portfolio groupings of the test assets are considered jointly rather than individually. A direct test of the joint null hypothesis may not be possible with standard methods when the total number of test assets grows large relative to the number of available time-series observations, since the estimate of the disturbance covariance matrix eventually becomes singular. The suggested residual bootstrap procedures based on combining the individual group p-values avoid this problem while controlling the overall significance level. Simulation and empirical results illustrate the usefulness of the joint mean-variance efficiency tests.

JEL classification: C12, C14, C15, G12
Bank classification: Econometric and statistical methods; Asset pricing; Financial markets

Résumé

Nous proposons des méthodes de bootstrap double pour tester l'hypothèse d'efficience moyenne-variance lorsque plusieurs groupes de portefeuilles d'actifs à tester sont examinés conjointement plutôt qu'individuellement. Il ne sera peut-être pas possible de tester directement l'hypothèse nulle conjointe au moyen de méthodes conventionnelles à partir du moment où il y a un nombre élevé d'actifs à tester par rapport au nombre de séries chronologiques disponibles, étant donné que l'estimation de la matrice de covariance des perturbations en vient ultimement à être singulière. Les procédures de bootstrap résiduel proposées, qui reposent sur la combinaison des différentes valeurs p des groupes individuels, permettent d'éviter ce problème tout en contrôlant le niveau de signification global. La simulation et les résultats empiriques mettent en lumière l'utilité des tests conjoints de l'hypothèse d'efficience moyenne-variance.

Classification JEL : C12, C14, C15, G12
Classification de la Banque : Méthodes économétriques et statistiques; Évaluation des actifs; Marchés financiers

1 Introduction

In the context of mean-variance analysis, a benchmark portfolio of assets is said to be efficient with respect to a given set of test assets if it is not possible to combine it with the test assets to obtain another portfolio with the same expected return as the benchmark portfolio, but a lower variance. With multiple benchmark portfolios, the question becomes whether some linear combination of them is efficient. The mean-variance efficiency hypothesis is a testable implication of the validity of linear factor asset pricing models, such as the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965), or more generally of the arbitrage pricing theory of Ross (1976); see Sentana (2009) for a survey of the econometrics of mean-variance efficiency tests.

A prominent way to assess the mean-variance efficiency hypothesis is with the test procedure of Gibbons et al. (1989) (GRS). This test takes the form of either a likelihood ratio or a system-wide F test conducted within a multivariate linear regression (MLR) model with as many equations as there are test assets in the cross-section. The exact distributional theory for the GRS test rests on the assumption that the MLR model disturbances are independent and identically distributed (i.i.d.) each period according to a multivariate normal distribution. Beaulieu et al. (2007) (BDK) extend the GRS test by developing a simulation-based procedure that allows for the possibility of non-Gaussian innovations. Another approach that also relaxes the GRS normality assumption is the residual bootstrap procedure of Chou and Zhou (2006) (CZ).

Any test procedure based on standard estimates of the MLR disturbance covariance matrix (e.g., GRS, BDK, CZ) requires that the size of the cross-section be less than the length of the time series in order to avoid singularities and hence be computable. A common practice is therefore to use portfolios rather than individual securities, whereby the test assets are sorted into portfolios according to some empirical characteristic such as the market value of the companies' equity and their book-to-market value. For instance, Gibbons et al. (1989) examine beta-sorted portfolios, industry-sorted portfolios and size-sorted portfolios. Shanken (1996) argues that creating portfolios also has the advantage of reducing the residual variance

and allowing the key regression parameters to be estimated more precisely.

Lewellen et al. (2010) suggest further that empirical tests of asset pricing models can be improved by expanding the set of test portfolios (beyond the commonly employed size and book-to-market portfolios) and using additional portfolios sorted by industry, beta, volatility, or factor loadings. They argue that a valid asset pricing model should be able to price all portfolios simultaneously. In this paper, we consider the problem of testing the mean-variance efficiency hypothesis when multiple portfolio sorts are grouped together and considered jointly rather than individually.[1] Observe that attempting a joint GRS, BDK or CZ test by taking all the portfolio groupings and stacking them into an MLR model may run into the singularity problem, since the expanded cross-section can exceed the length of the available time series. This issue will be even more pressing whenever the analysis is performed over short time periods, which is typically done to alleviate concerns about parameter stability.

The problem then consists of combining the tests for each portfolio grouping in a way that controls the overall level of the procedure. A difficulty in this situation is that even though the distribution of the individual test statistics might be known (e.g., under the GRS normality assumption), their joint distribution across portfolio groupings may be unknown or difficult to establish. In order to ensure that the overall significance level is no greater than, say, 5%, a smaller level must be used for each individual test. According to the well-known Bonferroni inequality, the individual levels should be set to 5% divided by the number of considered portfolio groupings. As this number grows, such Bonferroni-type adjustments can become far too conservative and lacking in power; see Savin (1984) for a survey discussion of these issues.

Westfall and Young (1993) explain in great detail that bootstrap methods can be used to solve multiple testing problems. Following these authors, we extend the CZ procedure and propose double bootstrap schemes à la Beran (1987, 1988) for controlling the overall significance level of mean-variance efficiency tests with multiple portfolio groupings. Specifically, the two methods we propose use statistics that combine the individual p-values from each

[1]For example, portfolios formed on size and book-to-market could be one grouping, while industry portfolios could be another.

portfolio grouping. The first method, which rests on the GRS normality assumption, takes the p-values from the marginal F distributions and then treats their combination like any other test statistic for the purpose of bootstrapping. The second (and more computationally expensive) method is entirely non-parametric in that a first level of bootstrapping is used to find the individual p-values in addition to the second level of bootstrapping for the combination of these p-values. Such double bootstrap schemes have been proposed by Godfrey (2005) to deal with multiple diagnostic tests in linear regression models; see also MacKinnon (2009) for a survey of these methods. Dufour et al. (2014) propose similar resampling-based methods for univariate regression models (with specified disturbance distributions) and apply them to serial dependence and predictability tests.

The current paper is organized as follows. In Section 2 we establish the statistical framework. We also describe the existing tests for a single portfolio grouping, including the Gibbons et al. (1989) and Chou and Zhou (2006) test procedures. We then discuss the problem of testing mean-variance efficiency with multiple portfolio groupings, and describe the proposed bootstrap methods. In Section 3 we illustrate the new tests by first comparing their relative performance in a simulation study and then by presenting the results of an empirical application. In Section 4 we offer some concluding remarks.

2 Framework and test procedures

We consider an investment universe comprising a risk-free asset, K benchmark portfolios and an additional set of N risky assets. At time t, the risk-free return is denoted r_{ft}, the returns on the benchmark portfolios are stacked in the $K \times 1$ vector $\mathbf{r}_{Kt}$, and, similarly, the returns on the other risky assets are stacked in the $N \times 1$ vector $\mathbf{r}_t$. Correspondingly, the time-t excess returns on the risky assets are denoted by $\mathbf{z}_{Kt} = \boldsymbol{r}_{Kt} - r_{ft}$ and $\mathbf{z}_t = \mathbf{r}_t - r_{ft}$.

Consider the following MLR model:

$$\mathbf{z}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{z}_{Kt} + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)'$ is an $N \times 1$ vector of intercepts (or *alphas*), $\boldsymbol{\beta}$ is an $N \times K$ matrix of

linear regression coefficients (or *betas*), and $\boldsymbol{\varepsilon}_t$ is an $N \times 1$ vector of model disturbances. These disturbances are such that $E[\boldsymbol{\varepsilon}_t \,|\, \mathbf{z}_{Kt}] = \mathbf{0}$ and $E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' \,|\, \mathbf{z}_{Kt}] = \boldsymbol{\Sigma}$, a non-singular covariance matrix. Jobson and Korkie (1982) show that if the usual expected return-beta representation $E[\mathbf{z}_t] = \boldsymbol{\beta} E[\mathbf{z}_{Kt}]$ holds, then some linear combination of the K benchmark portfolios is on the minimum-variance frontier. Therefore, a necessary condition for the efficiency of the K benchmark portfolios with respect to the N test assets is $H_0 : \boldsymbol{\alpha} = \mathbf{0}$ in the context of model (1). A direct test of H_0, however, may not be possible with standard methods when the size of the cross-section, N, is too large relative to the length of the time series, T. Indeed, the extant procedures described below to test mean-variance efficiency are based on the standard estimate of the covariance matrix of regression disturbances. As N grows relative to a fixed value of T, this matrix estimate eventually becomes singular and the usual tests can then no longer be computed.

A common practice in the application of mean-variance efficiency tests is thus to base them on portfolio groups in order to reduce the size of the cross-section of test assets. Dividing the securities into N_1 groups (such that $N_1 \leq T - K - 1$) solves the degrees-of-freedom problem with the original set of N test assets. As Shanken (1996) explains, portfolio diversification also has the potential effect of reducing the residual variances and increasing the precision with which the MLR alphas are estimated.[2]

2.1 Single portfolio grouping

Let $\mathbf{z}_{t,1}$ denote the $N_1 \times 1$ vector of returns obtained from grouping the test assets. According to model (1), these returns can be represented as

$$\mathbf{z}_{t,1} = \boldsymbol{\alpha}_1 + \boldsymbol{\beta}_1 \mathbf{z}_{Kt} + \boldsymbol{\varepsilon}_{t,1}, \tag{2}$$

[2]Shanken (1996) also discusses other motivations for the use of portfolio groupings. In particular, some stocks come and go over time and using portfolios allows the use of longer time series than would otherwise be possible. Forming portfolios also helps to prevent "survivorship biases," which result from the exclusion of failing stocks and thereby introduce an upward bias on the average sample return (Kothari et al., 1995). Also, portfolios formed by periodically ranking on some economic characteristic may be more likely to have constant betas compared to individual securities.

where $\boldsymbol{\alpha}_1$ and $\boldsymbol{\varepsilon}_{t,1}$ are both $N_1 \times 1$ vectors, and $\boldsymbol{\beta}_1$ is an $N_1 \times K$ matrix. The null hypothesis of interest corresponding to this grouping then becomes

$$H_0 : \boldsymbol{\alpha}_1 = \mathbf{0}. \tag{3}$$

Gibbons et al. (1989) propose a multivariate F test of H_0 in (3). Their test assumes that the vectors of disturbance terms $\boldsymbol{\varepsilon}_{t,1}$, $t = 1, ..., T$, in (2) are i.i.d. according to a multivariate normal distribution each period with mean zero and non-singular covariance matrix $\boldsymbol{\Sigma}_1$, conditional on $\mathbf{z}_{K1}, ..., \mathbf{z}_{KT}$.

Under normality, the methods of maximum likelihood and ordinary least squares (OLS) yield the same unconstrained estimates of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$:

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_1 &= \bar{\mathbf{z}}_1 - \hat{\boldsymbol{\beta}}_1 \bar{\mathbf{z}}_K, \\
\hat{\boldsymbol{\beta}}_1 &= \left[\sum_{t=1}^{T} (\mathbf{z}_{t,1} - \bar{\mathbf{z}}_1)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)' \right] \left[\sum_{t=1}^{T} (\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)' \right]^{-1},
\end{aligned}$$

where $\bar{\mathbf{z}}_1 = T^{-1} \sum_{t=1}^{T} \mathbf{z}_{t,1}$ and $\bar{\mathbf{z}}_K = T^{-1} \sum_{t=1}^{T} \mathbf{z}_{Kt}$. With $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\beta}}_1$ in hand, the unconstrained estimate of the disturbance covariance matrix is found as

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{z}_{t,1} - \hat{\boldsymbol{\alpha}}_1 - \hat{\boldsymbol{\beta}}_1 \mathbf{z}_{Kt} \right) \left(\mathbf{z}_{t,1} - \hat{\boldsymbol{\alpha}}_1 - \hat{\boldsymbol{\beta}}_1 \mathbf{z}_{Kt} \right)'. \tag{4}$$

The GRS test statistic for H_0 in (3) is

$$J_1 = \frac{(T - N_1 - K)}{N_1} \left[1 + \bar{\mathbf{z}}_K' \hat{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{z}}_K \right]^{-1} \hat{\boldsymbol{\alpha}}_1' \hat{\boldsymbol{\Sigma}}_1^{-1} \hat{\boldsymbol{\alpha}}_1, \tag{5}$$

where $\hat{\boldsymbol{\Omega}} = T^{-1} \sum_{t=1}^{T} (\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'$. Under the null hypothesis H_0, the statistic J_1 follows a central F distribution with N_1 degrees of freedom in the numerator and $(T - N_1 - K)$ degrees of freedom in the denominator. The statistic in (5) can also be written in the form of a likelihood ratio test (Campbell et al., 1997, Ch. 5). In either form, the GRS test is feasible only when $N_1 \leq T - K - 1$.

Beaulieu et al. (2007) extend the GRS test by developing an exact procedure based on likelihood ratios that allows for the possibility of non-Gaussian innovation distributions. Their framework assumes that the innovation distribution is either known or at least specified up to some unknown nuisance parameters. If normality is maintained, the BDK test becomes the Monte Carlo equivalent of the GRS test. Indeed, the BDK test procedure is akin to a parametric bootstrap with a finite-sample justification. When the assumed distribution involves unknown parameters (e.g., Student-t with unknown degrees of freedom), the BDK method proceeds by finding the maximal p-value over a confidence set for the intervening nuisance parameters. This confidence set is first established by numerically inverting a simulation-based goodness-of-fit test for the maintained distribution.

Chou and Zhou (2006) propose to use bootstrap methods to test mean-variance efficiency, avoiding the need to specify any distribution at all.[3] Of course, a non-parametric bootstrap is only asymptotically justified, but Chou and Zhou (2006) show that it works well even in small samples. The test statistic used in the CZ bootstrap procedure is the Wald ratio $W_1 = \hat{\boldsymbol{\alpha}}_1' \hat{\boldsymbol{\Sigma}}_1^{-1} \hat{\boldsymbol{\alpha}}_1$, which appears in the numerator of the GRS statistic in (5). Let $c_1 = \left[1 + \bar{\mathbf{z}}_K' \hat{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{z}}_K\right]^{-1} (T - N_1 - K)/N_1$ and observe that a bootstrap test based on W_1 is equivalent to one based on $J_1 = c_1 W_1$, since the term c_1 is constant under a fixed-regressor resampling scheme. Specifically, the CZ residual bootstrap procedure for the i.i.d. case considered here proceeds as follows:

1. Estimate the parameters of the MLR in (2) by OLS to obtain $\hat{\boldsymbol{\alpha}}_1$, $\hat{\boldsymbol{\beta}}_1$, and $\hat{\boldsymbol{\epsilon}}_{t,1} = \mathbf{z}_{t,1} - \hat{\boldsymbol{\alpha}}_1 - \hat{\boldsymbol{\beta}}_1 \mathbf{z}_{Kt}$ for $t = 1, ..., T$. Compute the Wald statistic as $W_1 = \hat{\boldsymbol{\alpha}}_1' \hat{\boldsymbol{\Sigma}}_1^{-1} \hat{\boldsymbol{\alpha}}_1$.

2. Estimate the MLR under the null hypothesis by setting the vector $\boldsymbol{\alpha}_1$ in (2) equal to zero and obtain

$$\tilde{\boldsymbol{\beta}}_1 = \left[\sum_{t=1}^{T} \mathbf{z}_{t,1} \mathbf{z}_{Kt}'\right] \left[\sum_{t=1}^{T} \mathbf{z}_{Kt} \mathbf{z}_{Kt}'\right]^{-1},$$

   where the tilde is used to distinguish this beta from its unconstrained counterpart, $\hat{\boldsymbol{\beta}}_1$.

3. For $i = 1, ..., B_1$, repeat the following steps:

[3] Hein and Westfall (2004) propose a similar residual bootstrap procedure to assess the significance of economic events for abnormal returns in the context of an MLR model.

(a) Generate bootstrap data according to $\mathbf{z}^*_{t,1,i} = \tilde{\boldsymbol{\beta}}_1 \mathbf{z}_{Kt} + \boldsymbol{\varepsilon}^*_{t,1,i}$ for $t = 1, ..., T$, where $\boldsymbol{\varepsilon}^*_{t,1,i}$ is drawn with replacement from $\{\hat{\boldsymbol{\varepsilon}}_{t,1}\}_{t=1}^T$.

(b) Apply OLS to the MLR model using the bootstrap data, thereby obtaining

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}^*_{1,i} &= \bar{\mathbf{z}}^*_{1,i} - \hat{\boldsymbol{\beta}}^*_{1,i} \bar{\mathbf{z}}_K, \\
\hat{\boldsymbol{\beta}}^*_{1,i} &= \left[\sum_{t=1}^{T} (\mathbf{z}^*_{t,1,i} - \bar{\mathbf{z}}^*_{1,i})(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)' \right] \left[\sum_{t=1}^{T} (\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)' \right]^{-1}, \\
\hat{\boldsymbol{\Sigma}}^*_{1,i} &= \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{z}^*_{t,1,i} - \hat{\boldsymbol{\alpha}}^*_{1,i} - \hat{\boldsymbol{\beta}}^*_{1,i} \mathbf{z}_{Kt} \right) \left(\mathbf{z}^*_{t,1,i} - \hat{\boldsymbol{\alpha}}^*_{1,i} - \hat{\boldsymbol{\beta}}^*_{1,i} \mathbf{z}_{Kt} \right)',
\end{aligned}$$

where $\bar{\mathbf{z}}^*_{1,i} = T^{-1} \sum_{t=1}^{T} \mathbf{z}^*_{t,1,i}$. Then compute the bootstrap Wald statistic as $W^*_{1,i} = \hat{\boldsymbol{\alpha}}^{*\prime}_{1,i} \hat{\boldsymbol{\Sigma}}^{*-1}_{1,i} \hat{\boldsymbol{\alpha}}^*_{1,i}$.

The null hypothesis H_0 in (3) should be rejected when the original statistic W_1 is in the upper tail. Using the simulated statistics $W^*_{1,1}, ..., W^*_{1,B_1}$, the bootstrap p-value is then simply

$$\hat{p}^* = \frac{1}{B_1} \sum_{i=1}^{B_1} \mathbb{I}[W^*_{1,i} > W_1], \tag{6}$$

where $\mathbb{I}[A]$ is the indicator function of event A, which is equal to 1 when A occurs and 0 otherwise. The decision rule consists of rejecting the null hypothesis when $\hat{p}^*$ is less than the nominal test level.

Observe that the bootstrap procedure uses the H_0-restricted estimate $\tilde{\boldsymbol{\beta}}_1$ when generating the artificial samples. This ensures that the bootstrap data are compatible with the null hypothesis. Also notice that the Chou and Zhou (2006) bootstrap method resamples the unrestricted residuals $\hat{\boldsymbol{\varepsilon}}_{1,1}, ..., \hat{\boldsymbol{\varepsilon}}_{T,1}$. By construction, these residuals have mean zero, thereby avoiding the need for centering. There is also an advantage in terms of power to using unrestricted rather than restricted residuals (MacKinnon, 2009).

In practical applications of mean-variance efficiency tests, we need to choose an appropriate number N_1 of test assets. It might seem natural to try to use as many as possible in

order to increase the probability of rejecting H_0 when it is false. Indeed, as the test asset universe expands it becomes more likely that non-zero pricing errors will be detected. However, as we have already mentioned, the choice of N_1 is restricted by T in order to keep the estimate of the disturbance covariance matrix in (4) from becoming singular, and the choice of T itself is often restricted owing to concerns about parameter stability. For instance, it is quite common to see studies where $T = 60$ monthly returns and N_1 is between 10 and 30.

The effect of increasing the number of test assets on test power is discussed in Gibbons et al. (1989), Campbell et al. (1997, p. 206), Sentana (2009), and Gungor and Luger (2013). When N_1 increases, there are in fact three effects that come into play: (i) the increase in the value of J_1's non-centrality parameter, which increases power, (ii) the increase in the number of degrees of freedom of the numerator, which decreases power, and (iii) the decrease in the number of degrees of freedom of the denominator due to the additional parameters that need to be estimated, which also decreases power. The additional caveat for the CZ resampling scheme is that the $N_1 \times N_1$ matrix $\hat{\boldsymbol{\Sigma}}^*_{1,i}$ will be singular and $W^*_{1,i}$ will not be defined with probability approaching 1 as N_1/T becomes large. For example, a bootstrap replication can sample the same N_1-vector $\boldsymbol{\varepsilon}^*_{t,1,i}$ too many times and in this case the rank of $\hat{\boldsymbol{\Sigma}}^*_{1,i}$ will be deficient.

To illustrate the net effect of increasing N_1 on the power of the GRS test and the CZ bootstrap test, we simulated model (2) with $K = 1$, where the returns on the single factor are random draws from the standard normal distribution. The elements of the independent disturbance vector were also drawn from the standard normal distribution, thereby ensuring the exactness of the GRS test. The elements of $\boldsymbol{\alpha}_1$ are generated randomly by drawing from a uniform distribution over $[-a, a]$, where we consider $a = 0.1$, 0.2 and 0.3. These values are well within the range of what we find with monthly stock returns. We set the sample size as $T = 60$ and we let the number of test assets N_1 range from 1 to 58.

Figure 1 shows the power of the GRS test (solid line) and the CZ bootstrap test computed with $B_1 = 1000$ (dashed line) as a function of N_1, where for any given value of N_1 the higher power curves are associated with a greater range $[-a, a]$. In line with the discussion in Gibbons et al. (1989), this figure clearly shows the power of the GRS test given this

specification rising as N_1 increases up to about one half of T, and then decreasing beyond that point. The results in Campbell et al. (1997, Table 5.2) show several other alternatives against which the power of the GRS test declines as N_1 increases. It is important to note that the choice of N_1 and T is somewhat arbitrary in practice, since there are no general results on how to devise an optimal multivariate test. Figure 1 also shows that the GRS and CZ tests have similar power when N_1 is between 1 and 5, but as more test assets are included, the power of the CZ bootstrap test peaks around $N_1 = 10$ and then falls to zero much sooner than the GRS test.

In addition to choosing N_1 and T, we must also select the assets used in the test procedure. A common practice is to group returns according to the ranked value of certain observable characteristics (e.g., industry, beta, size, book-to-market value, momentum) that are likely to offer a big spread in expected return deviations and boost the chances of rejecting the null hypothesis when the benchmark portfolios are not efficient. In the next section, we present two methods for testing mean-variance efficiency when multiple portfolio groupings are considered.[4]

2.2 Multiple portfolio groupings

Suppose there are G possible ways of dividing the test assets into groups, yielding the $N_g \times 1$ vectors $\mathbf{z}_{t,g}$, for $g = 1, ..., G$. These groups could differ in their number of included assets and/or selection of assets. By extension of (2), the groupings can be represented by the simultaneous equations

$$\mathbf{z}_{t,g} = \boldsymbol{\alpha}_g + \boldsymbol{\beta}_g \mathbf{z}_{Kt} + \boldsymbol{\varepsilon}_{t,g}, \quad g = 1, ..., G, \tag{7}$$

where we now have $\boldsymbol{\alpha}_g$ and $\boldsymbol{\varepsilon}_{t,g}$ as $N_g \times 1$ vectors, and $\boldsymbol{\beta}_g$ as an $N_g \times K$ matrix. Mean-variance efficiency implies the truth of $H_{0,g} : \boldsymbol{\alpha}_g = \mathbf{0}$, for all g. The joint null hypothesis of interest then becomes

$$H_0 : \text{ the hypotheses } H_{0,1}, ..., H_{0,G} \text{ are all true,} \tag{8}$$

[4]Expanding the set of test assets beyond the usual size and book-to-market portfolios is in fact the first prescription offered by Lewellen et al. (2010) to improve (cross-sectional) asset pricing tests.

which we wish to test in a way that keeps under control the overall probability of rejecting mean-variance efficiency when it actually holds.

Define the matrices $\mathbf{Z}_g = [\mathbf{z}_{1,g}, ..., \mathbf{z}_{T,g}]'$, $g = 1, ..., G$, and $\mathbf{X} = [\boldsymbol{\iota}_T, \mathbf{Z}_K]$, where $\boldsymbol{\iota}_T$ is a T-vector of ones and $\mathbf{Z}_K = [\mathbf{Z}_{K1}, ..., \mathbf{Z}_{KT}]'$. The models in (7) can then be written in the stacked MLR form:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \tag{9}$$

where $\mathbf{Y} = [\mathbf{Z}_1, ..., \mathbf{Z}_G]$ is a $T \times (N_1 + ... + N_G)$ matrix, $\mathbf{X}$ is a $T \times (K + 1)$ matrix of regressors, and $\mathbf{U} = [\mathbf{U}_1, ..., \mathbf{U}_G]$ is the $T \times (N_1 + ... + N_G)$ matrix of model disturbances defined with $\mathbf{U}_g = [\boldsymbol{\epsilon}_{1,g}, ..., \boldsymbol{\epsilon}_{T,g}]'$, $g = 1, ..., G$. The parameters are collected in $\mathbf{B} = [\mathbf{a}, \mathbf{b}]'$, a $(K + 1) \times (N_1 + ... + N_G)$ matrix, where $\mathbf{a} = [\boldsymbol{\alpha}_1', ..., \boldsymbol{\alpha}_G']'$ and $\mathbf{b} = [\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_G']'$. The system OLS estimates and residuals are given as usual by

$$\begin{aligned} \hat{\mathbf{B}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \\ \hat{\mathbf{U}} &= \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}, \end{aligned}$$

and it is well known that these are identical to what would be obtained if we applied OLS to each group separately before stacking. The disturbance covariance matrix estimate is computed as

$$\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{U}}'\hat{\mathbf{U}}/T,$$

but this matrix is singular when $(N_1 + ... + N_G) > T - K - 1$, meaning that parametric methods (like the GRS and BDK tests) cannot be applied to (9) for testing H_0 directly. So even though non-parametric test procedures are generally less powerful than parametric ones, taking a non-parametric route is the only option we have available for "large $\sum N_g$, small T" situations.[5]

Let us maintain for the moment the GRS assumption that the disturbance terms $\epsilon_{t,g}$,

[5]This point can also be seen in Affleck-Graves and McDonald (1990), who deal with a large number of test assets by using analogs to the GRS statistic computed with alternative covariance matrix estimators. Specifically, they consider an estimator based on the maximum entropy method of Theil and Laitinen (1980) and another one that restricts the covariance matrix to be diagonal. The distribution of the resulting statistics is then obtained via a residual bootstrap method.

$t = 1, ..., T$, in (7) are i.i.d. according to a multivariate normal distribution each period with mean zero and non-singular covariance matrix $\mathbf{\Sigma}_g$, conditional on the benchmark portfolio excess returns $\mathbf{z}_{K1}, ..., \mathbf{z}_{KT}$. Under this assumption and the joint H_0 in (8), the marginal GRS statistics J_g, $g = 1, ..., G$, each follow an F distribution with N_g degrees of freedom in the numerator and $(T - N_g - K)$ degrees of freedom in the denominator. Denote the corresponding marginal p-values by $p_g = 1 - F_{N_g, T-N_g-K}(J_g)$, where $F_{N_g, T-N_g-K}$ is the cumulative distribution function of the appropriate null distribution.

Following Dufour et al. (2014), we consider two methods of combining the individual p-values. The first one rejects H_0 when at least one of the individual p-values is sufficiently small. Specifically, if we define

$$p_{\min} = \min\{p_1, ..., p_G\} \text{ and } S_{\min} = 1 - p_{\min},$$

then we reject H_0 when $p_{\min}$ is small, or, equivalently, when $S_{\min}$ is large. The intuition here is that the null hypothesis should be rejected if at least one of the individual p-values is significant. The second combination method we consider is based on the product of the individual p-values:

$$p_\times = \prod_{g=1}^{G} p_g \text{ and } S_\times = 1 - p_\times,$$

which may provide more information about departures from H_0 compared to using only the minimum p-value.[6] To streamline the presentation, we next explain our inference methods with $S_{\min}$, and then we consider both $S_{\min}$ and $S_\times$ in our simulation study and empirical application.

We use bootstrap methods to estimate the distribution of $S_{\min}$ under H_0. Such resampling methods are necessary here in order to account for the dependence among the p-values. To see why, observe that the individual p-values are such that

$$p_g \sim U[0, 1] \text{ under } H_{0,g},$$

[6] We refer the reader to Folks (1984) for more on these and other test combination methods.

but only for each p-value taken one at a time. So even though the p-values $p_1, ..., p_G$ have identical marginal distributions under H_0, they need not be independent and may in fact have a very complex dependence structure. As Westfall and Young (1993) explain, bootstrap methods can be used to account for the correlation among the p-values and obtain a joint test of multiple hypotheses. These methods also avoid the need for Bonferroni-type adjustments, which quickly become far too conservative as G grows. See Godfrey (2005) and MacKinnon (2009) for related discussion and applications.

2.2.1 Bootstrap method I

The first method we propose exploits the Gaussian distributional assumption underlying the GRS test.[7] It should be noted, however, that even though this may seem like a stringent assumption, the GRS test is quite robust to typical departures from normality (Affleck-Graves and McDonald, 1989). The bootstrap method proceeds according to the following steps:

1. Estimate the parameters of the MLRs in (7) by OLS to obtain $\hat{\boldsymbol{\alpha}}_g$, $\hat{\boldsymbol{\beta}}_g$, and $\hat{\boldsymbol{\varepsilon}}_{t,g} = \mathbf{z}_{t,g} - \hat{\boldsymbol{\alpha}}_g - \hat{\boldsymbol{\beta}}_g \mathbf{z}_{Kt}$, for $t = 1, ..., T$ and $g = 1, ..., G$. Compute the GRS statistics as

$$J_g = c_g W_g = c_g \hat{\boldsymbol{\alpha}}_g' \hat{\boldsymbol{\Sigma}}_g^{-1} \hat{\boldsymbol{\alpha}}_g, \quad g = 1, ..., G,$$

   where $c_g = \left[1 + \bar{\mathbf{z}}_K' \hat{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{z}}_K\right]^{-1} (T - N_g - K)/N_g$.

2. Estimate the MLRs under the null hypothesis to obtain $\tilde{\boldsymbol{\beta}}_g$, $g = 1, ..., G$.

3. Compute $S_{\min} = 1 - p_{\min}$, where $p_{\min} = \min\{p_1, ..., p_G\}$ with $p_g = 1 - F_{N_g, T-N_g-K}(J_g)$.

4. For $i = 1, ..., B_1$, repeat the following steps:

   (a) Generate bootstrap data according to $\mathbf{z}_{t,g,i}^* = \tilde{\boldsymbol{\beta}}_g \mathbf{z}_{Kt} + \boldsymbol{\varepsilon}_{t,g,i}^*$, for $t = 1, ..., T$ and $g = 1, ..., G$, where the time-t collection $\boldsymbol{\varepsilon}_{t,1,i}^*, ..., \boldsymbol{\varepsilon}_{t,G,i}^*$ is drawn with replacement from $\{\hat{\boldsymbol{\varepsilon}}_{t,1}, ..., \hat{\boldsymbol{\varepsilon}}_{t,G}\}_{t=1}^T$.

[7] Method I thus includes the GRS test procedure as a special case (when $G = 1$).

(b) For $g = 1, ..., G$, apply OLS to the corresponding MLR model using the bootstrap data, thereby obtaining

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}^*_{g,i} &= \bar{\mathbf{z}}^*_{g,i} - \hat{\boldsymbol{\beta}}^*_{g,i}\bar{\mathbf{z}}_K, \\
\hat{\boldsymbol{\beta}}^*_{g,i} &= \left[\sum_{t=1}^{T}(\mathbf{z}^*_{t,g,i} - \bar{\mathbf{z}}^*_{g,i})(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'\right]\left[\sum_{t=1}^{T}(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'\right]^{-1}, \\
\hat{\boldsymbol{\Sigma}}^*_{g,i} &= \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{z}^*_{t,g,i} - \hat{\boldsymbol{\alpha}}^*_{g,i} - \hat{\boldsymbol{\beta}}^*_{g,i}\mathbf{z}_{Kt}\right)\left(\mathbf{z}^*_{t,g,i} - \hat{\boldsymbol{\alpha}}^*_{g,i} - \hat{\boldsymbol{\beta}}^*_{g,i}\mathbf{z}_{Kt}\right)',
\end{aligned}$$

where $\bar{\mathbf{z}}^*_{g,i} = T^{-1}\sum_{t=1}^{T}\mathbf{z}^*_{t,g,i}$. Then compute the bootstrap GRS statistic as $J^*_{g,i} = c_g\hat{\boldsymbol{\alpha}}^{*\prime}_{g,i}\hat{\boldsymbol{\Sigma}}^{*-1}_{g,i}\hat{\boldsymbol{\alpha}}^*_{g,i}$ and the corresponding p-value $p^*_{g,i} = 1 - F_{N_g,T-N_g-K}(J^*_{g,i})$.

(c) Compute $S^*_{\min,i} = 1 - p^*_{\min,i}$, where $p^*_{\min,i} = \min\{p^*_{1,i}, ..., p^*_{G,i}\}$.

The bootstrap p-value of $S_{\min}$ is then simply given by

$$\hat{p}^* = \frac{1}{B_1}\sum_{i=1}^{B_1}\mathbb{I}[S^*_{\min,i} > S_{\min}],$$

with small values suggesting that at least one of the hypotheses appearing in (8) may not be true. The formal decision rule is to reject the joint null hypothesis if $\hat{p}^*$ is less than the nominal significance level. Note that the p-value transformation p_g of the test statistic J_g to a quantile of the $U[0,1]$ distribution corresponds to the *prepivoting* step in Beran (1988).

A very important remark about this method (and the next one) is that the bootstrap samples are generated by randomly drawing the *entire* time-t collection $\boldsymbol{\varepsilon}^*_{t,1,i}, ..., \boldsymbol{\varepsilon}^*_{t,G,i}$ from $\{\hat{\boldsymbol{\varepsilon}}_{t,1}, ..., \hat{\boldsymbol{\varepsilon}}_{t,G}\}_{t=1}^{T}$. Stated in terms of the $T \times (N_1 + ... + N_G)$ matrix of system residuals associated with the stacked MLR in (9), the bootstrap proceeds by drawing entire rows of $\hat{\mathbf{U}}$. This kind of block resampling is the key for controlling the joint test size, since it preserves the contemporaneous correlation structure among the residuals.

2.2.2 Bootstrap method II

If we believe the MLR model innovations depart markedly from normality, then we can hedge against the risk of a misleading inference by bootstrapping the individual p-values in addition to bootstrapping their combination. This double bootstrap procedure works as follows:

1. Estimate the parameters of the MLRs in (7) by OLS to obtain $\hat{\boldsymbol{\alpha}}_g$, $\hat{\boldsymbol{\beta}}_g$, and $\hat{\boldsymbol{\varepsilon}}_{t,g} = \mathbf{z}_{t,g} - \hat{\boldsymbol{\alpha}}_g - \hat{\boldsymbol{\beta}}_g \mathbf{z}_{Kt}$, for $t = 1, ..., T$ and $g = 1, ..., G$. Compute the Wald statistics as

$$W_g = \hat{\boldsymbol{\alpha}}_g' \hat{\boldsymbol{\Sigma}}_g^{-1} \hat{\boldsymbol{\alpha}}_g, \quad g = 1, ..., G.$$

2. Estimate the MLRs under the null hypothesis to obtain $\tilde{\boldsymbol{\beta}}_g$, $g = 1, ..., G$.

3. For $i = 1, ..., B_1$, repeat the following steps:

   (a) Generate bootstrap data according to $\mathbf{z}_{t,g,i}^* = \tilde{\boldsymbol{\beta}}_g \mathbf{z}_{Kt} + \boldsymbol{\varepsilon}_{t,g,i}^*$, for $t = 1, ..., T$ and $g = 1, ..., G$, where the time-t collection $\boldsymbol{\varepsilon}_{t,1,i}^*, ..., \boldsymbol{\varepsilon}_{t,G,i}^*$ is drawn with replacement from $\{\hat{\boldsymbol{\varepsilon}}_{t,1}, ..., \hat{\boldsymbol{\varepsilon}}_{t,G}\}_{t=1}^T$.

   (b) For $g = 1, ..., G$, apply OLS to the corresponding MLR model using the bootstrap data, thereby obtaining

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_{g,i}^* &= \bar{\mathbf{z}}_{g,i}^* - \hat{\boldsymbol{\beta}}_{g,i}^* \bar{\mathbf{z}}_K, \\
\hat{\boldsymbol{\beta}}_{g,i}^* &= \left[\sum_{t=1}^T (\mathbf{z}_{t,g,i}^* - \bar{\mathbf{z}}_{g,i}^*)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'\right] \left[\sum_{t=1}^T (\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'\right]^{-1}, \\
\hat{\boldsymbol{\Sigma}}_{g,i}^* &= \frac{1}{T} \sum_{t=1}^T \left(\mathbf{z}_{t,g,i}^* - \hat{\boldsymbol{\alpha}}_{g,i}^* - \hat{\boldsymbol{\beta}}_{g,i}^* \mathbf{z}_{Kt}\right) \left(\mathbf{z}_{t,g,i}^* - \hat{\boldsymbol{\alpha}}_{g,i}^* - \hat{\boldsymbol{\beta}}_{g,i}^* \mathbf{z}_{Kt}\right)',
\end{aligned}$$

   where $\bar{\mathbf{z}}_{g,i}^* = T^{-1} \sum_{t=1}^T \mathbf{z}_{t,g,i}^*$. Then compute the bootstrap Wald statistic as $W_{g,i}^* = \hat{\boldsymbol{\alpha}}_{g,i}^{*\prime} \hat{\boldsymbol{\Sigma}}_{g,i}^{*-1} \hat{\boldsymbol{\alpha}}_{g,i}^*$.

4. With the simulated statistics $W^*_{g,1}, ..., W^*_{g,B_1}$, compute the first-level bootstrap p-values as

$$\hat{p}^*_g = \frac{1}{B_1}\sum_{i=1}^{B_1} \mathbb{I}[W^*_{g,i} > W_g], \quad g = 1, ..., G.$$

5. Compute $\hat{S}^*_{\min} = 1 - \hat{p}^*_{\min}$, where $\hat{p}^*_{\min} = \min\{\hat{p}^*_1, ..., \hat{p}^*_G\}$.

6. For $i = 1, ..., B_1$, do the following steps:

   (a) For $j = 1, ..., B_2$, do the following steps:

      i. Generate second-level bootstrap data according to $\mathbf{z}^{**}_{t,g,j} = \tilde{\boldsymbol{\beta}}_g \mathbf{z}_{Kt} + \boldsymbol{\varepsilon}^{**}_{t,g,j}$, for $t = 1, ..., T$ and $g = 1, ..., G$, where, as before, the time-t collection $\boldsymbol{\varepsilon}^{**}_{t,1,j}, ..., \boldsymbol{\varepsilon}^{**}_{t,G,j}$ is drawn with replacement from $\{\hat{\boldsymbol{\varepsilon}}_{t,1}, ..., \hat{\boldsymbol{\varepsilon}}_{t,G}\}_{t=1}^T$.

      ii. For $g = 1, ..., G$, apply OLS to the corresponding MLR model using the second-level bootstrap data, thereby obtaining

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}^{**}_{g,j} &= \bar{\mathbf{z}}^{**}_{g,j} - \hat{\boldsymbol{\beta}}^{**}_{g,j}\bar{\mathbf{z}}_K, \\
\hat{\boldsymbol{\beta}}^{**}_{g,j} &= \left[\sum_{t=1}^{T}(\mathbf{z}^{**}_{t,g,j} - \bar{\mathbf{z}}^{**}_{g,j})(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'\right]\left[\sum_{t=1}^{T}(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'\right]^{-1}, \\
\hat{\boldsymbol{\Sigma}}^{**}_{g,j} &= \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{z}^{**}_{t,g,j} - \hat{\boldsymbol{\alpha}}^{**}_{g,j} - \hat{\boldsymbol{\beta}}^{**}_{g,j}\mathbf{z}_{Kt}\right)\left(\mathbf{z}^{**}_{t,g,j} - \hat{\boldsymbol{\alpha}}^{**}_{g,j} - \hat{\boldsymbol{\beta}}^{**}_{g,j}\mathbf{z}_{Kt}\right)',
\end{aligned}$$

      where $\bar{\mathbf{z}}^{**}_{g,j} = T^{-1}\sum_{t=1}^{T}\mathbf{z}^{**}_{t,g,j}$. Then compute the second-level bootstrap Wald statistic as $W^{**}_{g,j} = \hat{\boldsymbol{\alpha}}^{**\prime}_{g,j}\hat{\boldsymbol{\Sigma}}^{**-1}_{g,j}\hat{\boldsymbol{\alpha}}^{**}_{g,j}$.

   (b) With the simulated statistics $W^{**}_{g,1}, ..., W^{**}_{g,B_2}$, compute the second-level bootstrap p-values as

$$\hat{p}^{**}_{g,i} = \frac{1}{B_2}\sum_{j=1}^{B_2} \mathbb{I}[W^{**}_{g,j} > W^*_{g,i}], \quad g = 1, ..., G.$$

   (c) Compute $\hat{S}^{**}_{\min,i} = 1 - \hat{p}^{**}_{\min,i}$, where $\hat{p}^{**}_{\min,i} = \min\{\hat{p}^{**}_{1,i}, ..., \hat{p}^{**}_{G,i}\}$.

The final test criterion of this double bootstrap method is

$$\hat{p}^{**} = \frac{1}{B_1} \sum_{i=1}^{B_1} \mathbb{I}[\hat{S}^{**}_{\min,i} > \hat{S}^{*}_{\min}],$$

which is just the proportion of simulated second-level combination statistics greater than $\hat{S}^{*}_{\min}$, the first-level combination statistic computed from the actual data. Note that here the first- and second-level bootstrap samples are generated the same way, so the asymptotic justification of this method is the same as for an ordinary (single) bootstrap test (cf. Beran, 1988).

Observe also that the double bootstrap is computationally expensive, since we need to calculate a total of $G(1 + B_1 + B_1 B_2)$ Wald test statistics. MacKinnon (2009) notes that the computational cost of performing a double bootstrap test can be substantially reduced by utilizing a stopping rule. Specifically, the replications can be stopped following the rules (for double bootstrap tests and confidence intervals) developed by Nankervis (2003, 2005) and the same results can be obtained as if all bootstrap calculations were used.

3 Illustrations

We illustrate the usefulness of the proposed bootstrap tests by applying them to the CAPM, which is a commonly applied model, in theory and in practice, for analyzing the trade-off between risk and expected return. Sharpe (1964) and Lintner (1965) show that if investors hold mean-variance efficient portfolios, then, under certain additional conditions, the market portfolio will itself be mean-variance efficient. The CAPM beta is the regression coefficient of the asset return on the single factor and it measures the systematic risk or co-movement with the returns on the market portfolio. Accordingly, assets with higher betas should in equilibrium offer higher expected returns.

The mean-variance CAPM takes the MLR form in (1) with $K = 1$ and a broad market index typically serves as a proxy for the market portfolio. Here we specify the market factor as the excess returns on a value-weighted stock market index of all stocks listed on the

NYSE, AMEX, and NASDAQ markets. The test assets are monthly excess returns on 25 size and book-to-market, 30 industry, 10 momentum, and 10 equity-price ratio portfolios (75 portfolios in total) over a 50-year period from January 1964 to December 2013 (600 months). Finally, we use the one-month U.S. Treasury bill as the risk-free asset when forming excess returns over the sample period.[8] It is also quite common in the empirical finance literature to test asset pricing models over subperiods owing to concerns about parameter stability (Campbell et al., 1997, Ch. 5). We follow this practice here and also perform the mean-variance efficiency tests over 5- and 10-year subperiods.

3.1 Simulation results

Before presenting the results of the empirical application, we first shed some light on the performance of the (CZ and new) bootstrap inference methods using the GRS test procedure as the benchmark for comparison purposes.

The artificial data are generated according to the single-factor version of (1) where $\mathbf{z}_{Kt} = \mathbf{z}_{1t}$ is obtained by randomly sampling the actual market factor. We also use in the data-generating process the actual estimates, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$, obtained from the 75 test asset portfolios over the full sample period. Specifically, for a given value of N, we populate the $N \times 1$ vector $\boldsymbol{\beta}$ by drawing randomly with replacement the elements of $\hat{\boldsymbol{\beta}}$, and the first $N \times N$ submatrix of $\hat{\boldsymbol{\Sigma}}$ serves as $\boldsymbol{\Sigma}$. The model disturbances are then generated as $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, thereby mimicking the cross-sectional covariance structure found among the actual sample residuals. In this setting, the GRS procedure is the uniformly most powerful invariant test (Affleck-Graves and McDonald, 1990). To examine the effects of non-normalities, we also consider disturbances drawn from a multivariate t-distribution with covariance matrix $\boldsymbol{\Sigma}$ and degrees of freedom equal to 20 and then 6. When investigating the relative power of the tests, the mispricing values $\alpha_1, ..., \alpha_N$ that make up $\boldsymbol{\alpha}$ in (1) are randomly drawn from a uniform distribution under two alternative scenarios: (i) $\alpha_i \sim U[-0.30, 0.30]$; and (ii) $\alpha_i \sim U[-0.35, 0.35]$. We consider sample sizes $T = 60, 120$, which correspond to 5 and 10 years of monthly data, and we vary the number of test assets as $N = 10, 30, 60, 75$.

[8]The data are obtained from Ken French's website at Dartmouth College.

The application of the new test procedures requires a choice about how to group the test assets. This choice obviously has no effect on the level of the tests, but, as Figure 1 shows, it matters for their power. The empirical rejection rates are therefore reported for several values of G and N_g to examine the effects of grouping. At the nominal 5% level, Tables 1–4 report the empirical size and power of the GRS, CZ, and new bootstrap (Methods I and II) tests, the latter being performed with the $S_{\min}$ and $S_{\times}$ combination statistics. The simulation design in Table 4 matches our empirical application in terms of sample size T, number of test assets $N = 75$, and number of groupings G. The bootstrap methods are implemented with $B_1 = 1000$, $B_2 = 100$, and the empirical rejection rates are based on 1000 replications of each data-generating configuration. The main findings of the simulation study can be summarized as follows.

1. When the GRS test is applicable, its empirical size is seen to stay close to the stated 5% level, even when the model disturbances follow a t-distribution; see Affleck-Graves and McDonald (1989) for further discussion about the robustness of the GRS test. The empirical rejection rates of the CZ test quickly decline to zero and the test ceases to be applicable as the ratio N/T increases, owing to the singularity of the estimated disturbance covariance matrix in the bootstrap world. In line with Figure 1, we see the CZ rejection rates in Tables 1–3 going to zero under both the null and alternative when $N/T \geq 1/2$. Observe that there are no CZ test results in Table 4, where N/T exceeds $1/2$.

2. The overall rejection rates of the double bootstrap tests depend not only on N/T, but also on the number of groupings G. For a fixed value of N/T, the empirical size shrinks toward zero as G decreases. Obviously, as G gets closer to 1, the new tests behave more like the original CZ test under the null and alternative hypotheses. Recall that when $G = 1$, Methods I and II correspond to the GRS and CZ test procedures, respectively.

3. In order to maximize the power of the bootstrap tests, it generally appears that G should be increased as N/T increases. From Tables 1–3 we see that when $N/T \leq 1/4$ (Panels A, B and D), we should set $G = 1$, i.e. perform the CZ test. An exception

occurs under the heavy-tailed t_6 distribution in Table 3, Panel D, where the bootstrap tests with $G = 2$ do slightly better. But as N/T increases, the new tests based on groupings clearly deliver more power than the CZ test. Indeed, the best power in Tables 1–3 seems to be with $G = 3$ when $N/T = 1/2$ (Panels C and F), and with $G = 6$ when $N/T = 1$ (Panel E). This pattern continues in Panels A, C and E of Table 4, where $N/T = 1.25$, and the best power performance occurs with $G = 7$ (6 groups with 10 portfolios each and a group of 15 portfolios).

4. The best bootstrap test power performances (set in bold) compare quite favourably to those of the GRS test. From Table 4 we can see that the power of bootstrap Methods I and II is on par with that of the GRS procedure, and can even surpass it. Indeed, in Panel F when the alternative is $\alpha_i \sim U[-0.35, 0.35]$, the GRS test has power of 78%, while the new bootstrap methods have power attaining 90% and more. Furthermore, when $N \geq T$ in Panel E of Tables 1–3 and Panels A, C, and E of Table 4, the double bootstrap tests are the only ones available.

5. In Tables 1–4, the new bootstrap methods appear to perform somewhat better with $S_\times$ than with S_{min} at the values of G that maximize power. For instance, in Panel A of Table 4 under $\alpha_i \sim U[-0.35, 0.35]$, bootstrap Method I applied with $G = 7$ delivers power of 55.4% with $S_\times$, versus 47.4% with S_{min}. The $S_\times$ statistic is also favoured with Method II.

6. Comparing the power performances of the two double bootstrap methods, we see that the completely non-parametric one (Method II) is only slightly less powerful than Method I, whose first-level p-values rest on the GRS normality assumption. A notable exception occurs in Table 4, Panels A, C, and E with $G = 5$, where Method II appears to outperform Method I. As expected, all the tests suffer relative power losses as the tails of the disturbance distribution become heavier from normal to t_{20} to t_6, and gain in power as T increases.

7. When the model disturbances deviate from normality (Tables 2 and 3; and Table 4, Panels C–F), the underlying GRS p-values used in Method I are only approximate.

Nevertheless, we see that the bootstrap procedure works remarkably well at keeping the test size under control. This finding concurs with the robustness results in Affleck-Graves and McDonald (1989), and is in line with the theoretical properties of combined p-values established by Dufour et al. (2014) in a parametric bootstrap context.

3.2 Empirical results

The results of the empirical application are reported in Table 5, where the entries are the p-values of the mean-variance efficiency tests performed with the $N = 75$ test asset portfolios over the full 50-year sample period, as well as 5- and 10-year subperiods. The entries set in bold represent cases of significance at the 5% level. Observe that the GRS and CZ tests are not computable with five years of monthly data, since $N = 75 > T = 60$. The CZ test remains "na" (not applicable) even with T=120 in the 10-year subperiods due to the singularity problem. We apply the new bootstrap methods with four different portfolio groupings: (i) $G = 7$ (6 groups of 10 portfolios and a group of 15 portfolios); (ii) $G = 5$ (2 groups of 10 portfolios, 2 groups of 15 portfolios, and a group of 25 portfolios); (iii) $G = 4$ (3 groups of 20 portfolios and a group of 15 portfolios); and (iv) $G = 3$ groups of 25 portfolios each.

For the 50-year period, the implications of the CAPM are strongly rejected by all the tests with p-values of no more than 0.02. In the 5-year subperiods, the new bootstrap tests indicate, for the most part, non-rejections of the mean-variance efficiency hypothesis. We also see some disagreements among the bootstrap tests. For instance, during the period 1/94–1/98, the decision as to whether to reject the null depends on the portfolio grouping. In light of the power results in Tables 1–4, we would naturally be inclined to agree with the rejections suggested by the $G = 7$ groups, since $N/T = 1.25$ in this case. Over the 10-year subperiods, the GRS and bootstrap tests agree on far more rejections of the null hypothesis at the conventional significance level. These results suggest that the CAPM generally finds more support over shorter periods of time and tends to be incompatible with the data as the time span lengthens. Note that the wild fluctuation in bootstrap p-values already revealed by the 5-year subperiods is suggestive of temporal instabilities in the CAPM representation

of expected returns.

4 Conclusion

In this paper we have described how double bootstrap methods can be used to test the mean-variance efficiency hypothesis in the presence of multiple portfolio groupings. Under the null hypothesis, the MLR model intercepts should be zero no matter how the test assets are divided into groups. There are two ways we could test these joint restrictions. First, we may stack the portfolio groups into an MLR model with, say, $G \times N_g$ equations and proceed either with the F test of Gibbons et al. (1989) or the residual bootstrap method of Chou and Zhou (2006). The shortcoming of this "testing by stacking" approach is that the multivariate GRS and CZ tests may lose all their power or may not even be computable as $G \times N_g$ becomes large relative to T. This problem can be clearly seen in Figure 1. In comparison to the unconditional GRS test, the singularity problem appears much sooner in the conditional CZ bootstrap world.

Instead of testing by stacking, we proposed a "divide and conquer" approach, which proceeds by bootstrapping combinations of the individual p-values associated with each portfolio grouping. The individual p-values may be obtained from the marginal F distributions, if we assume that the MLR disturbances are (not too far from) normally distributed. We showed how these p-values can be combined (using either the minimum p-value or their product) into a single statistic, which is then treated like any other statistic for the purpose of bootstrapping. The second method we suggested uses a first round of bootstrapping to find the individual p-values in addition to the second layer of bootstrap replications used to get the p-value of the combined statistic. Of course, this second method is computationally more demanding than the first one, but it offers protection in situations where the marginal GRS p-values may be grossly incorrect. These double bootstrap methods account for the possibly complex dependence structure among the p-values and control the probability of rejecting the joint null hypothesis when mean-variance efficiency actually holds.

References

Affleck-Graves, J. and B. McDonald (1989). Nonnormalities and tests of asset pricing theories. *Journal of Finance 44*, 889–908.

Affleck-Graves, J. and B. McDonald (1990). Multivariate tests of asset pricing: the comparative power of alternative statistics. *Journal of Financial and Quantitative Analysis 25*, 163–185.

Beaulieu, M.-C., J.-M. Dufour, and L. Khalaf (2007). Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors. *Journal of Business and Economic Statistics 25*, 398–410.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika 74*, 457–468.

Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association 83*, 687–697.

Campbell, J., A. Lo, and A. MacKinlay (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Chou, P.-H. and G. Zhou (2006). Using bootstrap to test portfolio efficiency. *Annals of Economics and Finance 1*, 217–249.

Dufour, J.-M., L. Khalaf, and M. Voia (2014). Finite-sample resampling-based combined hypothesis tests, with applications to serial correlation and predictability. *Communications in Statistics – Simulation and Computation, forthcoming*.

Folks, J. (1984). Combination of independent tests. In P. Krishnaiah and P. Sen (Eds.), *Handbook of Statistics 4: Nonparametric Methods*, pp. 113–121. North-Holland, Amsterdam.

Gibbons, M., S. Ross, and J. Shanken (1989). A test of the efficiency of a given portfolio. *Econometrica 57*, 1121–1152.

Godfrey, L. (2005). Controlling the overall significance level of a battery of least squares diagnostic tests. *Oxford Bulletin of Economics and Statisitcs 67*, 263–279.

Gungor, S. and R. Luger (2013). Testing linear factor pricing models with large cross-sections: a distribution-free approach. *Journal of Business and Economic Statistics 31*, 66–77.

Hein, S. and P. Westfall (2004). Improving tests of abnormal returns by bootstrapping the multivariate regression model with event parameters. *Journal of Financial Econometrics 2*, 451–471.

Jobson, J. and B. Korkie (1982). Potential performance and tests of portfolio efficiency. *Journal of Financial Economics 10*, 433–466.

Kothari, S., J. Shanken, and R. Sloan (1995). Another look at the cross-section of expected stock returns. *Journal of Finance 50*, 185–224.

Lewellen, J., S. Nagel, and J. Shanken (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics 96*, 175–194.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics 47*, 13–37.

MacKinnon, J. (2009). Bootstrap hypothesis testing. In D. Belsley and J. Kontoghiorghes (Eds.), *Handbook of Computational Econometrics*, pp. 183–213. Wiley.

Nankervis, J. (2003). Stopping rules for double bootstrap tests. *Working Paper No. 03/14, Department of Accounting, Finance and Management, University of Essex*.

Nankervis, J. (2005). Computational algorithms for double bootstrap confidence intervals. *Computational Statistics and Data Analysis 49*, 461–475.

Ross, S. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*, 341–360.

Savin, N. (1984). Multiple hypothesis testing. In Z. Griliches and M. Intriligator (Eds.), *Handbook of Econometrics*, pp. 827–879. North-Holland, Amsterdam.

Sentana, E. (2009). The econometrics of mean-variance efficiency: a survey. *Econometrics Journal 12*, 65–101.

Shanken, J. (1996). Statistical methods in tests of portfolio efficiency: a synthesis. In G. Maddala and C. Rao (Eds.), *Handbook of Statistics, Vol. 14: Statistical Methods in Finance*, pp. 693–711. Elsevier Science B.V.

Sharpe, W. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance 19*, 425–442.

Theil, H. and K. Laitinen (1980). Singular moment matrices in applied econometrics. In P. Krishnaiah (Ed.), *Multivariate Analysis*, pp. 629–649. North-Holland, Amsterdam.

Westfall, P. and S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.

Table 1. Empirical size and power: normal disturbances, $N = 10, 30, 60$ test assets

	Size: $\alpha_i = 0$						Power: $\alpha_i \sim U[-0.3, 0.3]$						Power: $\alpha_i \sim U[-0.35, 0.35]$					
	GRS	CZ	Method I		Method II		GRS	CZ	Method I		Method II		GRS	CZ	Method I		Method II	
$G \times N_g$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$
Panel A: $N = 10, T = 60$																		
1 × 10	5.0	3.6					42.3	**36.0**					53.3	**49.0**				
2 × 5			3.8	4.0	3.5	3.8			29.4	30.0	29.2	29.7			41.5	41.3	41.0	40.0
5 × 2			5.5	5.1	5.1	4.8			19.4	11.0	17.6	10.6			25.7	15.6	24.9	13.4
Panel B: $N = 10, T = 120$																		
1 × 10	3.7	3.2					78.4	**76.8**					88.5	**87.4**				
2 × 5			3.5	3.5	3.7	4.0			66.9	66.4	65.3	64.9			79.9	80.3	78.0	79.6
5 × 2			3.5	4.3	3.5	4.3			44.2	31.1	44.0	25.8			58.6	42.8	56.6	37.0
Panel C: $N = 30, T = 60$																		
1 × 30	6.4	0.0					51.8	0.0					70.4	0.2				
2 × 15			1.2	1.0	1.4	1.1			32.4	35.8	31.2	34.9			48.0	55.2	46.5	54.9
3 × 10			2.6	2.3	3.1	2.5			**39.2**	**44.6**	**37.2**	**43.7**			**57.9**	**62.1**	**54.8**	**61.3**
5 × 6			4.7	3.4	4.6	3.1			34.7	37.0	34.2	34.2			50.8	55.6	48.3	52.7
6 × 5			4.0	3.4	4.2	3.5			34.6	35.5	33.1	32.9			50.3	52.1	47.6	48.4
10 × 3			4.5	4.0	4.9	4.3			26.8	18.9	26.6	18.0			38.2	26.7	36.2	25.1
15 × 2			5.3	5.6	5.0	6.1			23.9	10.2	22.7	11.2			33.4	12.5	31.0	13.7
Panel D: $N = 30, T = 120$																		
1 × 30	4.0	1.5					98.6	**95.2**					99.4	**98.9**				
2 × 15			3.5	3.1	3.1	2.8			94.9	96.5	93.3	95.5			98.2	99.0	98.0	98.9
3 × 10			4.1	4.3	3.4	3.9			92.4	95.2	91.1	94.3			97.6	99.2	96.9	99.1
5 × 6			4.4	5.3	5.2	4.9			84.8	90.7	83.1	89.1			95.0	97.8	93.9	96.9
6 × 5			4.5	5.2	5.0	5.1			82.7	88.8	80.7	85.8			93.8	96.8	92.5	95.5
10 × 3			5.3	5.5	4.8	4.9			70.5	49.2	66.7	47.9			85.9	56.1	83.5	54.3
15 × 2			5.3	5.3	4.8	5.1			62.2	19.1	58.1	15.8			77.4	19.7	73.5	16.7
Panel E: $N = 60, T = 60$																		
1 × 60	na	na					na	na					na	na				
3 × 20			0.7	0.2	0.7	0.3			11.0	9.7	10.5	8.9			19.9	20.1	21.2	20.8
5 × 12			1.3	0.9	1.6	0.8			26.9	29.1	26.8	28.1			43.2	47.8	41.8	44.5
6 × 10			2.7	1.3	2.8	1.4			**26.5**	**29.7**	**25.9**	**28.1**			**44.5**	**49.7**	**41.8**	**45.3**
10 × 6			3.4	2.8	3.7	3.1			29.2	23.9	27.9	23.7			41.5	36.7	39.0	34.4
15 × 4			4.3	4.8	3.8	4.2			25.6	13.4	22.3	11.5			36.6	14.8	33.6	13.1
Panel F: $N = 60, T = 120$																		
1 × 60	4.2	0.0					98.8	2.2					99.9	9.4				
2 × 30			0.8	1.0	1.7	1.1			94.5	95.3	93.4	93.4			98.8	99.7	98.7	99.3
3 × 20			2.9	1.6	2.9	1.7			**96.1**	**97.9**	**95.0**	**96.6**			**99.5**	**99.8**	**99.1**	**99.7**
5 × 12			3.0	3.3	3.0	3.0			92.7	97.2	92.3	95.9			99.6	99.9	98.7	99.4
6 × 10			3.5	2.9	3.6	2.9			92.0	96.6	90.5	95.0			98.0	99.8	98.0	99.4
10 × 6			3.9	2.8	3.6	3.2			82.5	57.7	79.3	62.6			94.7	60.5	93.8	65.8
15 × 4			4.7	5.0	4.7	5.7			73.0	17.7	69.5	18.6			86.3	17.7	83.3	18.9

Notes: This table reports the empirical size and power (in percentages) of the GRS, CZ, and the proposed (bootstrap methods I and II) tests with a total of $N = 10, 30, 60$ test assets over $T = 60, 120$ time periods. The nominal level is 5% and the results are based on 1000 replications. The entries in bold indicate the portfolio groupings that tend to maximize the power of the bootstrap tests and the abbreviation "na" stands for not applicable.

Table 2. Empirical size and power: t_{20} disturbances, $N = 10, 30, 60$ test assets

	Size: $\alpha_i = 0$						Power: $\alpha_i \sim U[-0.3, 0.3]$						Power: $\alpha_i \sim U[-0.35, 0.35]$					
	GRS	CZ	Method I		Method II		GRS	CZ	Method I		Method II		GRS	CZ	Method I		Method II	
$G \times N_g$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$
Panel A: $N = 10, T = 60$																		
1 × 10	4.2	3.2					37.7	**30.2**					51.8	**43.7**				
2 × 5			4.7	3.7	4.9	3.8			29.1	27.3	28.3	26.4			40.1	37.6	37.5	36.7
5 × 2			4.7	4.6	4.6	4.0			19.7	11.3	18.7	10.5			25.6	14.4	24.5	13.9
Panel B: $N = 10, T = 120$																		
1 × 10	5.0	4.8					76.4	**74.6**					86.7	**85.6**				
2 × 5			4.6	4.6	5.0	4.7			64.2	63.4	62.9	62.6			76.1	76.4	75.0	75.0
5 × 2			4.7	4.5	4.8	4.5			42.5	27.7	40.9	24.2			54.6	39.2	52.9	34.3
Panel C: $N = 30, T = 60$																		
1 × 30	4.9	0.0					53.8	0.1					69.2	0.1				
2 × 15			1.5	0.7	1.4	1.2			29.1	33.6	28.1	32.0			45.2	50.2	43.0	50.5
3 × 10			2.9	1.9	2.7	2.1			**36.0**	**41.9**	**33.2**	**39.4**			**51.1**	**59.7**	**47.2**	**58.4**
5 × 6			4.0	3.3	4.1	3.2			32.7	33.7	30.4	33.1			46.6	52.6	45.9	50.1
6 × 5			4.8	4.2	4.7	4.0			31.8	32.8	31.9	31.3			47.3	48.0	44.6	45.9
10 × 3			4.7	4.6	5.1	4.7			25.4	16.7	26.7	18.9			36.0	23.4	36.1	25.4
15 × 2			4.2	4.2	4.3	5.7			21.9	9.4	21.7	11.5			31.1	11.1	29.2	13.5
Panel D: $N = 30, T = 120$																		
1 × 30	4.6	1.1					98.4	**93.8**					99.8	**98.7**				
2 × 15			3.7	2.9	3.4	2.8			91.2	94.5	89.3	93.8			98.1	99.3	97.4	99.0
3 × 10			3.9	3.6	3.5	3.2			89.0	93.9	87.2	93.0			96.5	99.1	95.8	98.8
5 × 6			4.7	4.6	4.3	4.4			81.9	86.8	78.6	85.3			92.5	95.9	91.6	95.0
6 × 5			4.9	4.5	4.1	3.9			77.7	83.4	74.6	80.7			91.6	94.7	89.8	93.0
10 × 3			5.1	5.3	4.9	5.0			62.2	44.4	59.3	43.4			80.1	53.2	76.2	51.2
15 × 2			4.7	4.2	5.2	5.3			55.2	16.4	52.7	17.3			70.8	17.2	68.9	18.3
Panel E: $N = 60, T = 60$																		
1 × 60	na	na					na	na					na	na				
3 × 20			0.2	0.0	0.1	0.0			6.5	5.6	7.1	5.0			14.3	14.4	14.3	14.7
5 × 12			1.5	0.6	1.5	0.7			20.7	20.3	21.0	20.6			37.5	40.0	35.5	36.6
6 × 10			1.6	1.1	2.2	1.0			**23.5**	**23.2**	**23.6**	**23.0**			**39.1**	**40.9**	**37.5**	**39.9**
10 × 6			2.9	3.4	3.1	2.2			23.0	18.0	21.3	18.8			35.6	27.1	34.3	28.8
15 × 4			3.5	3.7	3.6	5.1			20.6	11.2	20.8	13.4			30.7	13.3	28.4	15.1
Panel F: $N = 60, T = 120$																		
1 × 60	4.7	0.0					96.3	0.7					99.4	3.9				
2 × 30			0.4	0.4	0.4	0.3			87.9	89.9	86.6	88.1			96.9	98.0	96.1	97.0
3 × 20			1.3	1.1	1.9	1.3			**91.3**	**94.0**	**89.6**	**93.0**			**97.9**	**98.9**	**97.8**	**98.8**
5 × 12			4.1	1.7	4.2	2.0			88.7	92.7	86.7	91.1			97.1	99.0	96.9	98.5
6 × 10			3.8	2.3	3.8	2.5			85.2	92.3	85.1	90.4			96.0	98.5	94.4	97.4
10 × 6			4.8	3.4	4.8	4.4			77.7	55.8	74.0	59.8			91.3	60.5	89.0	65.9
15 × 4			5.8	5.1	6.0	5.6			62.6	18.0	60.6	18.6			81.2	18.3	78.4	19.1

Notes: This table reports the empirical size and power (in percentages) of the GRS, CZ, and the proposed (bootstrap methods I and II) tests with a total of $N = 10, 30, 60$ test assets over $T = 60, 120$ time periods. The nominal level is 5% and the results are based on 1000 replications. The entries in bold indicate the portfolio groupings that tend to maximize the power of the bootstrap tests and the abbreviation "na" stands for not applicable.

Table 3. Empirical size and power: t_6 disturbances, $N = 10, 30, 60$ test assets

	Size: $\alpha_i = 0$						Power: $\alpha_i \sim U[-0.3, 0.3]$						Power: $\alpha_i \sim U[-0.35, 0.35]$					
	GRS	CZ	Method I		Method II		GRS	CZ	Method I		Method II		GRS	CZ	Method I		Method II	
$G \times N_g$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$
Panel A: $N = 10, T = 60$																		
1 × 10	4.5	2.4					30.2	**21.8**					41.3	**29.2**				
2 × 5			3.8	4.2	3.4	3.5			20.2	18.5	19.6	19.0			28.5	28.1	28.5	26.6
5 × 2			4.7	5.3	4.0	5.3			16.2	10.3	16.5	10.1			22.5	13.5	21.1	13.4
Panel B: $N = 10, T = 120$																		
1 × 10	5.9	4.6					62.8	**60.2**					77.1	**73.7**				
2 × 5			5.5	4.8	4.8	4.8			48.9	48.9	48.6	48.2			64.0	62.2	61.4	61.7
5 × 2			3.7	3.7	4.0	3.6			29.1	18.6	27.6	15.9			38.7	27.5	37.4	22.6
Panel C: $N = 30, T = 60$																		
1 × 30	4.7	0.0					46.2	0.0					61.5	0.0				
2 × 15			0.5	0.5	0.5	0.3			13.6	15.4	13.1	14.0			24.4	28.5	23.3	26.0
3 × 10			1.4	1.3	1.5	1.5			**20.5**	**22.8**	**21.2**	**22.4**			**32.8**	**36.2**	**31.6**	**35.7**
5 × 6			2.5	2.1	3.2	2.7			20.1	19.7	21.6	19.9			30.0	31.3	29.1	28.8
6 × 5			2.3	2.4	2.8	2.7			20.2	18.9	18.6	18.2			29.5	30.0	28.9	28.2
10 × 3			3.5	3.4	3.5	3.6			16.9	12.5	16.6	12.0			23.9	16.6	23.3	15.7
15 × 2			4.4	3.9	4.7	4.7			14.0	8.5	14.3	9.0			20.9	9.9	20.2	10.5
Panel D: $N = 30, T = 120$																		
1 × 30	4.0	0.4					95.2	78.5					99.3	92.8				
2 × 15			1.8	1.7	2.0	1.8			**81.2**	**86.0**	**78.1**	**84.4**			**93.4**	**95.5**	**91.4**	**94.2**
3 × 10			2.6	2.6	2.9	2.4			78.4	84.3	75.7	82.0			91.5	94.6	90.1	94.0
5 × 6			3.8	3.9	3.8	3.9			68.5	73.9	66.4	69.3			84.1	88.7	80.9	86.0
6 × 5			3.6	4.3	3.1	4.1			64.9	69.5	62.2	65.7			80.5	86.9	79.7	84.2
10 × 3			4.0	4.4	3.8	4.9			49.5	34.2	48.0	35.6			67.4	45.3	63.7	45.8
15 × 2			4.7	4.7	4.7	6.3			41.7	14.5	40.4	17.0			58.0	16.1	55.3	18.7
Panel E: $N = 60, T = 60$																		
1 × 60	na	na					na	na					na	na				
3 × 20			0.0	0.0	0.0	0.0			3.4	1.9	3.4	2.2			7.9	5.2	7.0	6.0
5 × 12			0.6	0.6	1.1	0.7			13.1	12.4	14.0	12.8			22.8	22.7	22.1	22.6
6 × 10			1.2	0.7	1.1	0.7			**15.4**	**14.8**	**15.7**	**13.8**			**24.3**	**24.1**	**24.2**	**23.4**
10 × 6			2.8	1.9	2.8	2.0			18.5	11.6	18.2	14.5			27.7	17.5	25.6	20.5
15 × 4			3.5	3.9	3.6	4.7			16.4	8.5	17.0	10.7			24.0	9.9	23.3	13.1
Panel F: $N = 60, T = 120$																		
1 × 60	5.9	0.0					87.7	0.1					90.8	0.6				
2 × 30			0.0	0.0	0.3	0.2			67.6	68.2	66.0	66.4			87.5	87.1	86.0	86.0
3 × 20			0.7	0.9	0.9	0.9			**75.0**	**78.0**	**72.3**	**74.7**			**90.4**	**91.5**	**88.4**	**90.0**
5 × 12			2.3	1.9	2.3	1.9			71.6	77.4	70.8	74.1			89.0	91.3	27.3	89.2
6 × 10			2.8	2.3	3.1	2.9			69.2	74.6	68.0	71.7			84.8	89.6	84.5	86.6
10 × 6			3.9	3.9	3.9	3.9			56.0	42.0	53.1	45.4			76.1	50.8	72.2	56.4
15 × 4			3.8	4.7	3.9	4.0			45.9	15.5	44.2	15.0			65.6	16.6	61.0	16.1

Notes: This table reports the empirical size and power (in percentages) of the GRS, CZ, and the proposed (bootstrap methods I and II) tests with a total of $N = 10, 30, 60$ test assets over $T = 60, 120$ time periods. The nominal level is 5% and the results are based on 1000 replications. The entries in bold indicate the portfolio groupings that tend to maximize the power of the bootstrap tests and the abbreviation "na" stands for not applicable.

Table 4. Empirical size and power: $N = 75$ test assets

	Size: $\alpha_i = 0$						Power: $\alpha_i \sim U[-0.3, 0.3]$						Power: $\alpha_i \sim U[-0.35, 0.35]$					
	GRS	CZ	Method I		Method II		GRS	CZ	Method I		Method II		GRS	CZ	Method I		Method II	
G			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$			$S_{\min}$	$S_{\times}$	$S_{\min}$	$S_{\times}$
Panel A: normal disturbances, $T = 60$																		
1	na	na					na	na					na	na				
3			0.0	0.0	0.0	0.0			1.3	0.8	1.3	0.9			3.4	3.1	3.4	3.0
4			0.2	0.2	0.2	0.2			11.9	11.9	12.5	12.7			23.0	27.5	23.3	26.3
5			0.2	0.2	1.4	0.8			7.8	13.0	21.9	20.5			16.0	28.9	33.6	35.2
7			1.4	1.3	2.4	2.0			**29.3**	**34.2**	**28.5**	**31.2**			**47.4**	**55.4**	**43.7**	**52.1**
Panel B: normal disturbances, $T = 120$																		
1	3.7	na					98.2	na					100.0	na				
3			1.9	1.7	2.1	2.5			95.8	98.2	95.2	97.6			99.7	100.0	99.2	100.0
4			2.9	2.6	2.9	3.2			**96.2**	**98.7**	**95.2**	**98.4**			**99.9**	**99.9**	**99.8**	**99.8**
5			2.9	3.0	3.6	3.2			95.0	98.0	93.4	96.5			99.4	100.0	98.8	99.9
7			5.1	4.3	5.0	4.5			94.9	97.8	95.5	92.5			99.0	99.9	98.5	99.4
Panel C: t_{20} disturbances, $T = 60$																		
1	na	na					na	na					na	na				
3			0.0	0.0	0.0	0.0			0.5	0.5	0.5	0.4			1.5	2.1	1.5	1.5
4			0.1	0.1	0.6	0.2			7.9	8.3	9.2	9.2			16.1	19.1	18.8	20.9
5			0.1	0.1	2.0	0.5			4.9	9.1	16.6	15.6			9.3	20.9	26.9	27.6
7			1.9	0.9	1.9	1.2			**23.4**	**26.2**	**23.1**	**24.7**			**37.9**	**43.6**	**36.3**	**41.6**
Panel D: t_{20} disturbances, $T = 120$																		
1	5.1	na					97.3	na					99.5	na				
3			1.4	0.7	1.5	1.2			91.3	96.6	89.0	94.5			98.8	99.3	97.6	99.1
4			2.2	1.4	2.1	1.5			**93.9**	**97.3**	**92.5**	**96.0**			**98.7**	**99.6**	**97.7**	**99.3**
5			2.8	2.1	3.3	2.2			90.1	95.1	88.7	93.0			98.0	99.3	97.2	98.8
7			4.2	2.6	4.6	2.5			88.4	95.4	86.1	93.0			97.4	99.3	95.8	97.8
Panel E: t_6 disturbances, $T = 60$																		
1	na	na					na	na					na	na				
3			0.0	0.0	0.0	0.0			0.0	0.2	0.3	0.0			0.3	0.7	0.8	0.2
4			0.1	0.1	0.1	0.1			2.4	2.8	3.8	2.9			6.7	6.8	7.2	7.3
5			0.2	1.0	0.2	0.0			1.4	3.4	9.1	5.7			3.5	8.8	15.0	11.9
7			0.8	0.9	1.4	0.9			**12.4**	**12.9**	**15.1**	**12.2**			**21.3**	**22.9**	**21.9**	**22.2**
Panel F: t_6 disturbances, $T = 120$																		
1	7.5	na					76.5	na					78.1	na				
3			0.9	0.6	0.6	0.6			73.3	77.1	70.8	74.9			90.2	93.0	87.9	90.9
4			2.0	1.4	1.6	1.2			**76.5**	**81.8**	**73.6**	**79.0**			**92.3**	**94.4**	**89.8**	**93.8**
5			2.0	1.3	2.6	1.5			71.5	79.4	71.2	76.3			90.2	94.0	87.6	91.9
7			3.9	2.0	3.2	2.2			69.5	78.8	67.5	76.3			87.3	91.9	85.2	91.6

Notes: This table reports the empirical size and power (in percentages) of the GRS, CZ, and the proposed (bootstrap methods I and II) tests with a total of $N = 75$ test assets over $T = 60, 120$ time periods. The number of portfolio groupings is: (i) $G = 1$ (1 group of 75 portfolios); (ii) $G = 3$ groups of 25 portfolios each; (iii) $G = 4$ (3 groups of 20 portfolios and a group of 15 portfolios); (iv) $G = 5$ (2 groups of 10 portfolios, 2 groups of 15 portfolios, and a group of 25 portfolios); and (v) $G = 7$ (6 groups of 10 portfolios and a group of 15 portfolios). The nominal level is 5% and the results are based on 1000 replications. The entries in bold indicate the portfolio groupings that tend to maximize the power of the bootstrap tests and the abbreviation "na" stands for not applicable.

Table 5. Mean-variance efficiency test results

	GRS	CZ	Method I								Method II							
			$S_{\min}$				$S_{\times}$				$S_{\min}$				$S_{\times}$			
			$G=7$	5	4	3	7	5	4	3	7	5	4	3	7	5	4	3
50-year period																		
1/64–12/13	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**	**0.02**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
5-year subperiods																		
1/64–12/68	na	na	0.07	0.26	0.17	0.42	0.06	0.09	0.13	0.26	0.12	0.16	0.10	0.42	**0.05**	0.09	0.11	0.21
1/69–12/73	na	na	0.25	0.34	0.52	0.55	0.17	0.27	0.51	0.43	0.28	0.15	0.60	0.61	0.17	0.20	0.51	0.49
1/74–12/78	na	na	0.41	0.53	0.67	0.76	0.27	0.29	0.62	0.67	0.37	0.39	0.76	0.82	0.27	0.22	0.62	0.70
1/79–12/83	na	na	0.47	0.59	0.65	0.75	0.50	0.44	0.53	0.64	0.55	0.44	0.68	0.77	0.49	0.40	0.57	0.68
1/84–12/88	na	na	**0.00**	**0.03**	**0.03**	0.17	**0.00**	**0.01**	**0.00**	0.14	**0.02**	**0.01**	**0.02**	0.13	**0.02**	**0.05**	**0.04**	0.09
1/89–12/93	na	na	0.09	0.31	0.14	0.49	0.15	0.30	0.14	0.43	0.15	0.18	0.19	0.49	0.24	0.42	0.22	0.45
1/94–12/98	na	na	**0.00**	0.13	**0.02**	0.25	**0.05**	0.08	0.07	0.15	**0.01**	0.24	**0.02**	0.21	**0.03**	**0.05**	0.12	0.14
1/99–12/03	na	na	0.12	0.37	0.38	0.48	0.31	0.45	0.47	0.65	0.07	0.56	0.44	0.63	0.34	0.53	0.50	0.71
1/04–12/08	na	na	0.07	0.37	0.31	0.49	0.06	0.26	0.23	0.53	**0.04**	0.21	0.23	0.46	**0.04**	0.21	0.20	0.53
1/09–12/13	na	na	0.18	0.50	0.38	0.65	0.11	0.39	0.37	0.59	0.16	0.56	0.35	0.48	0.13	0.35	0.32	0.38
10-year subperiods																		
1/64–12/73	**0.02**	na	**0.02**	**0.01**	0.07	**0.01**	**0.00**	**0.01**	**0.02**	**0.00**	0.07	**0.04**	0.06	**0.03**	**0.02**	**0.02**	**0.02**	**0.02**
1/74–12/83	0.11	na	0.09	0.10	0.29	0.20	0.06	0.07	0.14	0.23	0.09	0.12	0.25	0.19	**0.04**	0.08	0.10	0.21
1/84–12/93	**0.00**	na	**0.00**	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**	**0.01**	**0.02**	**0.01**	**0.01**	**0.01**	**0.02**	**0.01**	**0.01**
1/94–12/03	**0.00**	na	**0.00**	**0.01**	**0.00**	**0.01**	**0.03**	**0.03**	**0.01**	**0.04**	**0.01**	**0.02**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**	0.07
1/04–12/13	**0.01**	na	0.20	0.06	0.18	0.08	0.07	0.13	0.11	**0.04**	0.22	0.11	0.10	0.07	0.07	0.15	0.13	**0.02**

Notes: The table entries are the p-values of the GRS, CZ, and the proposed (bootstrap methods I and II) tests. The test assets comprise 25 size and book-to-market, 30 industry, 10 momentum, and 10 earnings-price ratio portfolios. The proposed tests are applied to these $N = 75$ portfolios by dividing them into $G = 7, 5, 4, 3$ portfolio groupings. The market portfolio return is the value-weighted return on NYSE, AMEX, and NASDAQ stocks and the risk-free rate is the 1-month Treasury bill rate. The entries set in bold indicate significant cases at the 5% level and the abbreviation "na" stands for not applicable.

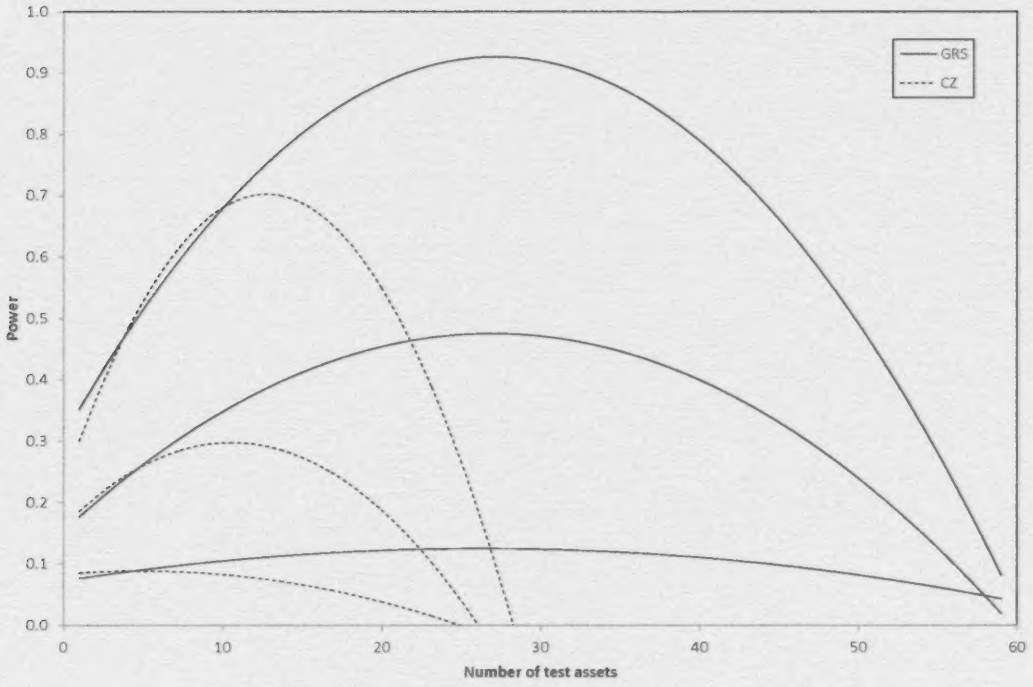

Figure 1. This figure plots the power of the GRS and CZ tests as a function of the number of included test assets. The returns are generated from model (2) with normally distributed innovations. The sample size is $T = 60$ and the number of test assets N_1 ranges from 1 to 58. The tests are performed at the nominal 0.05 level and the higher power curves are associated with greater expected return deviations.